Journal of Informetrics 13 (2019) 887–900

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

How does collaboration affect researchers' positions in co-authorship networks?

Xiangjie Kong[a], Mengyi Mao[a], Huizhen Jiang[b], Shuo Yu[a], Liangtian Wan[a,*]

[a] Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China
[b] Peking Union Medical College Hospital, Beijing 100730, China

ARTICLE INFO

Article history:
Received 19 October 2018
Received in revised form 30 July 2019
Accepted 31 July 2019
Available online 26 August 2019

Keywords:
Social analytics
Triadic closure
Bacon number
Generalized friendship paradox

ABSTRACT

Collaboration usually has a positive effect on researchers' productivity: researchers have become increasingly collaborative, according to recent studies. Numerous studies have focused on enhancing research collaboration by recommendation technology and measuring the influence of researchers. However, few studies have investigated the effect of collaboration on the position of a researcher in the research social network. In this paper, we explore the relationships between collaboration and influence by social analytical methods, which are pertinent to analyzing the network structure and individual traits. We evaluate three aspects of the researchers' influence: friendship paradox validation, social circle, and structure of a researcher's ego network. Furthermore, the "six degrees of Bacon number" theory, generalized friendship paradox, and triadic closure theory are introduced to support our analysis. Experimental results show that collaboration can help researchers increase their influence to some extent.

1. Introduction

We live in a world of networks with numerous nodes and connections. Social network analysis can help us understand some social phenomena and behaviors (Schoenebeck, 2013; Zhong, Fan, Zhu, & Yang, 2013), and helps to reveal and understand the network structure and features of individual nodes (Staiano et al., 2012). For example, analysis of online social networks such as Facebook and Twitter shows users' personality and a tendency to associate with similar users (Liu, Venkatanathan, Goncalves, Karapanos, & Kostakos, 2014; Weitzel, Quaresma, & de Oliveira, 2012). Nowadays, the grand challenges that humans confront include rapidly changing human society and population structure, global crises, and increasing rates of crime. To solve these problems, it is necessary and meaningful for social scientists to analyze the structure of society and patterns based on massive amounts of data (Conte et al., 2012). For example, Liu et al. (Liu et al. 2018) studied the detailed structure of citation networks, in terms of the evolution of topics in artificial intelligence (AI), to improve the comprehension of the development of human society. Computational social science aims to model social problems quantitatively, understand social systems (Conte et al., 2012), and explore some mechanisms behind these phenomena. As a result, it promotes more studies of application services, such as recommender systems (Hsiao, Kulesza, & Hero, 2014; Kong, Mao, Wang, Liu, & Xu, 2018; Rafailidis & Crestani, 2016).

* Corresponding author.
E-mail address: wanliangtian@dlut.edu.cn (L. Wan).

https://doi.org/10.1016/j.joi.2019.07.005
1751-1577/

Collaboration has been a vitally important behavioral phenomenon for social scientists to analyze social patterns, whether in education (Barr & Gunawardena, 2012), development teams (Ghobadi, 2015), or research (Coccia & Wang, 2016). Social analytical studies of co-authorship networks give a quantitative description of collaboration and may be pertinent for mining the essence of co-authorship networks (Kong et al., 2016, 2017; Wang et al., 2017; Xia, Wang, Bekele, & Liu, 2017). Collaboration is a very important aspect of research activity. In particular, persistent academic collaborations promote the progress and development of the whole academia. Studies of research collaboration reveal the facilitating relationships between research achievements and researchers' cooperation (Amjad et al., 2017; Mavin, 2015; Zhang, Bu, & Ding, 2016). A two-stage least squares analysis proposed by Lee et al. (Lee & Bozeman, 2005) confirmed that collaboration has a positive effect on researchers' productivity. Triadic closure can be applied to find more collaborations and understand transitivity in a co-authorship network. There is another social analytical theory, the *friendship paradox*, which states that one's friends have more friends than oneself on average. The friendship paradox has been confirmed to hold for more than 98% of Twitter users (Hodas, Kooti, & Lerman, 2013). The friendship paradox can be extended to some research characteristics, in which case it is called the *generalized friendship paradox*. One's coauthors are likely to be more prominent than oneself, on average, with more publications, more citations, and more collaborators (Eom & Jo, 2014; Grund, 2014). We have noticed that a small amount of elite researchers, who become influential in their social circles, can break the law of the friendship paradox. This raises the question, how does collaboration benefit researchers in co-authorship networks? In this work, we will analyze co-authorship networks by social analytical methods to investigate this phenomenon.

The goal of this paper is to investigate what an influential researcher looks like in a co-authorship network. An influential researcher usually performs prolifically and has a higher rank in his/her social circle, attracting more people to collaborate (Zhang, Bu, Ding, & Xu, 2017; Zhang et al., 2016). His/her social distance from famous researchers is shorter than that of others. In this work, we mainly analyze the influence that collaboration can have on research, based on the DBLP dataset. We focus on three aspects. First, we explore the friendship paradox based on the researchers' rank. Second, from the viewpoint of the sociability of researchers, we study the benefits of collaboration on enhancing researchers' social circles. Finally, we introduce the Bacon number (Backstrom, Boldi, Rosa, Ugander, & Vigna, 2012), which measures the distance from one actor to Kevin Bacon. We can measure the social distance between an arbitrary researcher and a famous researcher, and explore the importance of the social circle of a given scholar. We evaluate the scope and dynamic change of the Bacon Number in a co-authorship network. We also consider two network centrality metrics, i.e. clustering coefficient and network density to measure the importance of scholars in the network. According to this analysis, we explore the effect of collaboration on the position of a researcher in a co-authorship network.

2. Related work

2.1. Generalized friendship paradox

The friendship paradox originates from Feld (Feld, 1991). It is a sociological phenomenon, which follows the structural properties of social networks, that most people have fewer friends than their friends do on average. However, it indicates that most people think they are more popular than their friends (Zuckerman & Jost, 2001). Studies of the social network of Facebook (Ugander, Karrer, Backstrom, & Marlow, 2011) found that the number of friends on Facebook was 190, on average, but the average number of friends of their friends was 635. The same was found for Twitter: your followers and those that you follow have more followers than you and follow more people than you on average (Hodas et al., 2013). This phenomenon also exists in social networks offline (Eom & Jo, 2014; Grund, 2014). The friendship paradox is actually just a simple mathematical result, but there are also many fantastic practical meanings, which may not be the original intention of their discoverers.

The generalized friendship paradox is an extension of the friendship paradox, applied to node characteristics other than degree. Eom and Jo (Eom & Jo, 2014) introduced the definition of the generalized friendship paradox and found that one's collaborators are likely to be more prominent on average, with more publications, more citations, and more collaborators in the scientific co-authorship network. Fotouhi et al. (Fotouhi, Momeni, & Rabbat, 2015) considered a network growth model that is a preferential attachment scheme, to assess how the generalized friendship paradox affects the distribution of node qualities. Momeni et al. (Momeni & Rabbat, 2016) studied the generalized friendship paradox in the context of online social networks, through the lens of measures of activity and influence. Lerman et al. (Lerman, Yan, & Wu, 2016) studied the "majority illusion" to explore the relationships between network influence and the friendship paradox. Benevenuto et al. (Benevenuto, Laender, & Alves, 2016) explored the h-index paradox in co-authorship networks. However, fewer studies have focused on the impact of research collaboration on researchers influenced by the generalized friendship paradox, and how to break the generalized friendship paradox with the dynamics of co-authorship networks.

2.2. Triadic closure

Most analyses of social networks are based on the static network, which is a snapshot of a social network at a certain time. The analyses study the relationships between nodes, and the relationships between nodes and edges, at a fixed time. However, it is more meaningful to analyze the process and rules for the evolution of social networks over time. Triadic closure provides a good angle from which to analyze networks dynamically.

The concept of triadic closure was proposed by Simmel in 1908 and developed by Granovetter (Granovetter, 1977), who added the principle of strong ties and weak ties. Triadic closure explains a basic principle in the evolution of a social network: if two individuals do not know each other but they have a common friend, the probability of those two individuals becoming friends in the future will be greater. For three nodes A, B, and C, if there is a strong tie between A and B as well as between A and C, there is a high probability of a weak or strong tie between B and C. To quantify this tendency, the terms "clustering" and "community detection" occur in the network literature, and several algorithms and metrics have been proposed to detect the phenomenon. Watts and Strogatz (Watts & Strogatz, 1998) introduced the clustering coefficient, which measures the density of a node's ego network. Barabási et al. (Barabási et al., 2002) demonstrated the likelihood of an author's collaborators working together in the future to explain this clustering in co-authorship networks. Newman (Newman, 2001) proposed a new metric, the global clustering coefficient. The more common friends two authors have, the higher the probability that they can be friends in the future, which explains link formation mechanisms in co-authorship networks.

Following Newman's method to interpret clustering coefficients for triadic closure in co-authorship networks, substantial work has been conducted. Newman's metric is directly applied to a one-mode network instead of a two-mode network, which can overestimate the tendency of triadic closure in a co-authorship network (particularly if a paper has more than three authors). To improve the applicability of clustering coefficients, Opsahl (Opsahl, 2013) captured triadic closure into two-mode networks as closed four-paths, to exclude false positives for triadic closure. Based on Opsahl's work, Kim et al. (Kim & Diesner, 2017) proposed an over-time version of Opsahl's metric, which is suitable for inferring edge formation in evolving co-authorship networks.

2.3. Scientific collaboration

The trend of increasing scientific collaboration has been confirmed in many studies (Amjad et al., 2017; Barabási et al., 2002; Coccia & Wang, 2016; Lee & Bozeman, 2005; Zhang et al., 2017). Among these studies of scientific collaboration, the patterns of collaboration (Barabási et al., 2002; Coccia & Wang, 2016) and the importance of collaboration are two key research topics. For evaluating the impact of researchers (Dunaiski, Geldenhuys, & Visser, 2018) and papers, and the value of collaboration, the most common methods use the number of researchers' papers and citations. The centrality of scientists in a network can influence the number of citations; conversely, citations can influence how researchers are evaluated and can increase the visibility of their future work (Sarigol, Pfitzner, Scholtes, Garas, & Schweitzer, 2014). Other influencing factors, like the number of direct scientific partners and the career age of the researchers, have been applied to assess researchers, based on their positions and roles in the co-authorship network (Ebadi & Schiffauerova, 2015; Wang et al., 2017). Moreover, we can also evaluate a researcher by other metrics, such as a journal's impact factor, the rank of a venue, the gender of the researcher (Thelwall, 2018), the researcher's h-index, or even the rank of the institute (van Dijk, Manor, & Carey, 2014). Concerning the rank of collaborators or journals, the PageRank algorithm (Page, Brin, Motwani, & Winograd, 1999) can be applied to quantify the impact of researchers (Nykl, Campr, & Jezek, 2015) using different weighting functions. Therefore, we applied PageRank to evaluate research rank scores in Section 3.1. For collaboration behavior analysis, network centrality metrics, such as closeness centrality and eigenvector centrality, are also common metrics. The review of related literature reveals that many studies have focused on the recommendation or prediction of influential researchers, or measuring the rank and impact of researchers, but we want to explore the mechanism responsible for collaboration influence.

3. Methods

To explore the relationships between collaboration and researchers' positions in the co-authorship network, some social analytical methods on co-authorship networks have been adopted. We introduce them in this section.

3.1. Friendship paradox validation

According to the generalized friendship paradox theory, there are only a small number of researchers who can perform more influentially than their friends do, on average. To validate the generalized friendship paradox theory, we can explore the mechanism from the angle of the researchers' rank. First, we need to quantify researchers' achievements and measure researchers' rank. There is a well-known network-based method, PageRank (Page et al., 1999), which can rank all nodes in the network according to their importance and degree of popularity. In the case of co-authorship networks, PageRank can be used to measure collaboration value to some extent (Xia, Chen, Wang, Li, & Yang, 2014). In this work, we model co-authorship networks according to this rule: researchers are regarded as nodes, and an edge will be added between two researchers if they have coauthored at least one paper.

Academic Rank (*AR*): We define the symbol *AR* to represent the academic rank score, which measures the rank of a researcher in the co-authorship network. The computational process of *AR* is shown in Eq. (1), which looks like a generalized PageRank, and the computation is similar to PageRank.

$$AR^{t+1} = \alpha \cdot \mathbf{S} \cdot AR^{t} + (1-\alpha) \cdot q, \tag{1}$$

where $(1-\alpha)$ represents the probability that the next neighbor node is randomly selected and matrix $\mathbf{S}$ denotes the relationship between researchers in the co-authorship network.

Besides the *AR* score, the number of a researcher's collaborators can reflect the researcher's popularity. The achievement of researchers can also be represented by the number of their publications. A prolific researcher usually produces a large amount of publications. We calculate three metrics: *AR* score, number of collaborators, and number of publications. The average values of these metrics are also calculated for the researcher's collaborators. Through these analyses, we can explore the friendship paradox in the co-authorship network, to some extent.

In practice, we observe a dynamic change in these three metrics as the number of collaborators increases. The statistics in Section 4 will show how a researcher becomes influential in the co-authorship network.

3.2. Social circle analysis

Next, we want to measure the achievements of researchers from the social circle. The scale of researchers' social circles can approximately reflect their research influence. Generally, influential researchers perform better in cooperation (Amjad et al., 2017; Lee & Bozeman, 2005). Having a well-connected and wide-ranging social circle is an obvious characteristic of an influential researcher. Moreover, the importance of a researcher's social circle can also reflect the influence of the researcher. To measure whether collaboration affects a researcher's position in the network, we evaluate the four metrics below while changing the researchers' social circles by varying the number of collaborators. The theory of triadic closure describes the fact that two people may get to know each other if they have the same friends (Opsahl, 2013). This reveals the phenomenon that collaboration can increase the probability of acquiring new collaborators in social activities, which will enhance a researcher's social circle and make it more well-connected. These new collaborators occupy a very important part of the social circle. After a new collaboration, all members of the collaborator's social circle will become the researcher's potential collaborators. The number of researchers at a two-step social distance can reflect researchers' potential social circles. We first propose our assumption according to the triadic closure theory. We assume that a new co-authorship between a researcher and collaborators of his/her collaborators is caused by his/her collaborators. We then give a clear method of calculating these metrics.

Number of new collaborators: The new collaborators for a certain researcher refer to those collaborators who are brought by existing collaboration relationships. It can also be regarded as the researcher gets to know these new collaborators from his/her collaborators. For example, if *A* collaborates with *B*'s collaborator *C* after *A* collaborated with *B*, then *C* becomes *A*'s new collaborator. However, if *A* builds the collaboration relationship with *D*, then *D* cannot be regarded as a new collaborator if *D* has not collaborated with any collaborator of *A*.

Proportion of new collaborators: The proportion of new collaborators is used to evaluate the ration of a researcher's newly increased collaborators. We calculate this metric for each researcher. For example, there are three collaborators *B*, *C*, and *D* of researcher *A*. If *B* and *D* have already collaborated with *A* before the first collaboration between *A* and *C*, then *C* is *A*'s new collaborator. That is, *C* is the new comer of *A* compared with *B* and *D*. The proportion of new collaborators for researcher *A* is 1/3.

Number of potential collaborators: It is the sum of the "collaborators of collaborators", that is, the researchers at a two-step social distance with the potential for collaboration. We assume that the collaborators of a researcher' collaborators have the potential to become direct collaborators of the researcher. For example, after researchers *A* and *B* build a co-authorship, all of *B*'s collaborators who are not *A*'s collaborators are regarded as *A*'s potential collaborators.

Bacon Number[1] : This is usually used to evaluate the social distance of other actors from Kevin Bacon, based on the original "Erdos number" (Castro & Grossman, 1999) in the mathematical co-authorship network. In this work, we use the Bacon Number to analyze the co-authorship network for the first time. Fig. 1 shows a simple co-authorship network under the theoretical framework of six degrees of separation (Leskovec & Horvitz, 2008). We assume that there are seven researchers including a famous powerful researcher, Bacon. Researcher *a* needs only one step to reach Bacon, so his/her Bacon Number is 1, while researcher *f* needs six steps to reach Bacon, bridged by researchers *e*, *d*, *c*, *b*, and *a*, so his/her Bacon Number is 6. If Bacon is defined as the core researcher in his/her research field, we can say that researcher *a* is more influential than *f*, because *a* is closer to the key node than *f* in the co-authorship network. To calculate the Bacon Number for each researcher, we first need to discover Bacon, the most influential and powerful researcher in the co-authorship network. We adopt the *AR* score from Section 3.1 to measure researchers' importance and popularity in their social circles. We then regard the top researcher, according to *AR* score, as "Bacon".

We will measure the variation of these four metrics as the number of collaborators increases. This can show the benefits that collaboration brings to researchers in their social circles.

3.3. Structure of researchers' ego network analysis

In this section, we analyze the relationships between the influence of researchers and the number of their collaborators through each researcher's ego network structure. First, we explore the structure of researchers' ego networks from the perspective of centrality, including eigenvector centrality, betweenness centrality, and closeness centrality.

[1] https://en.wikipedia.org/wiki/Six_Degrees_of_Kevin_Bacon#Undefined_Bacon_numbers

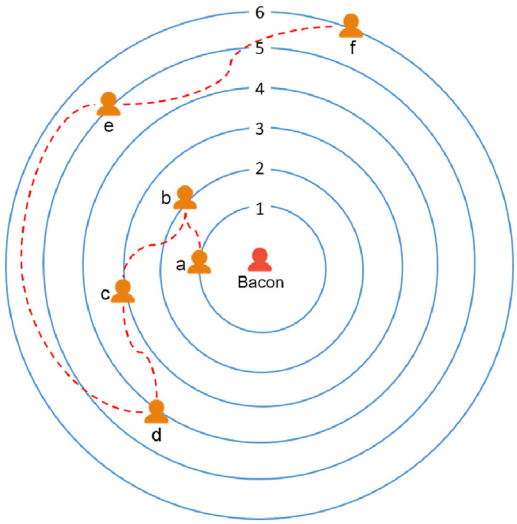

Fig. 1. A simple social network illustrating the Bacon Number.

Eigenvector centrality [2] is a metric to measure the influence of a node in a network. The centrality score of a node is determined by the scores of neighbor nodes. All nodes in the network are assigned relative scores based on the following standard:

$$x_v = \frac{1}{\lambda}\sum_{t \in G} a_{v,t} x_t. \tag{2}$$

For a given graph G, if node v is linked to node t, $a_{v,t} = 1$, otherwise $a_{v,t} = 0$. λ represents the greatest eigenvalue of the adjacency matrix $A = (a_{v,t})$. According to Eq. (2), the score of a node is the sum of its adjacent nodes' scores. This means that a connection to a high-scored node has a greater contribution to the target node than a connection to a low-scored node.

Betweenness centrality [3] is a metric based on the shortest paths in a network. The betweenness centrality of a node is the number of shortest paths between all pairs of nodes in a network that pass through the node. The shortest paths minimize the total number of edges (for an unweighted network) or the weights of the edges (for a weighted network). Betweenness centrality, which is widely used in network theory, represents the degree of dependence between the nodes. The betweenness centrality of a node v is shown in Eq. (3):

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{3}$$

where σ_{st} is the number of all shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through node v. If a node has a higher betweenness centrality, it is more important and influential in the network and gains more information from the network. Briefly, the betweenness centrality of the node v is its importance when acting as a bridge.

Closeness centrality [4] measures whether a node is at the core of the network. The main idea of closeness centrality is that a node will be seldom manipulated by other nodes and will interact with all other nodes more easily if it is central. Closeness centrality is calculated in Eq. (4), which refers to the sum of shortest path lengths from the node to all other node.

$$C_c(v) = \frac{n-1}{\sum_{i=1}^{n} d(v,i)}, \tag{4}$$

where $d(v, i)$ is the distance between node v and node i, and n is the number of nodes in the network. It is noteworthy that the network must be connected when we calculate the closeness centrality.

We also analyze the influence and importance of researchers in ego networks, including two metrics: clustering coefficient and network density. Clustering coefficient is to measure the compactness of nodes in a network that tend to cluster together. The nodes tend to gather together with relatively high edge density, which is greater than the average probability of establishing edges between two nodes randomly. There are two kinds of clustering coefficients: global and local. The global clustering coefficient refers to the overall amount of clustering in a network, while the local clustering coefficient

[2] https://en.wikipedia.org/wiki/Eigenvector_centrality
[3] https://en.wikipedia.org/wiki/Betweenness_centrality
[4] https://en.wikipedia.org/wiki/Closeness_centrality

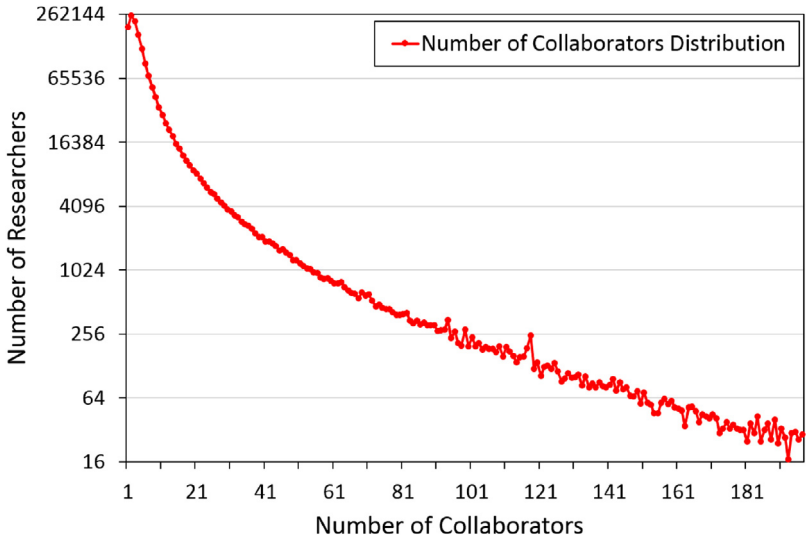

Fig. 2. Distribution of researcher degree from dataset. The y-axis has a log scale.

refers to the aggregation of a single node. In this paper, we consider the local clustering coefficient, which is defined in Eq. (5):

$$CC_v = \frac{2n}{k(k-1)}, \tag{5}$$

where k is the degree of node v and n is the number of edges between all neighbor nodes of node v. Network density is mainly used to measure the compactness of edges among nodes in a network. If the interactions between nodes in a network are more frequent, the density of this network is relatively high. For a network G with n nodes and l edges, the definition of network density is:

$$d(G) = \frac{2l}{n(n-1)}. \tag{6}$$

4. Results and discussions

We conducted extensive implementation of the methods described in the previous section. All of the statistics and analysis were based on the DBLP dataset, a computer science bibliography website hosted at the University of Trier in Germany that contains bibliographic information for papers, such as the title and authors. In this section, we describe the detailed implementation and discuss the analysis results.

4.1. Data

DBLP provides open bibliographic information on major computer science journals and proceedings. The entire DBLP dataset can be found at `http://dblp.uni-trier.de/xml/`. It has indexed 3,272,991 papers and 1,752,443 authors up to June 2016. When we built the co-authorship network based on DBLP, we considered a researcher as a node, a coauthor relationship between two researchers as an edge, and found the maximal connected component subgraphs, to build the unweighted co-authorship network as G. Finally, we extracted 1,511,153 authors and 7,247,543 coauthor relationships of G. All metrics mentioned in Section 3 (*AR* score, number of publications, number of collaborators, number of new collaborators, proportion of new collaborators, number of potential collaborators, Bacon Number, eigenvector centrality, betweenness centrality, closeness centrality, clustering coefficient, and network density) were calculated based on G for each researcher.

To present our statistical results, we divided the researchers into different categories according to the number of collaborators. For example, we considered those researchers having one collaborator as one category and those researchers with 100 collaborators as another category. We selected 100 researchers randomly for each category (for example, we randomly selected 100 researchers from the 197784 researchers with one collaborator; see Fig. 2), and then we calculated the average metrics of these 100 researchers and recorded the minimum and maximum for the following analysis. In particular, the number of researchers who have a large number of collaborators is less than 100, so all researchers in those categories were included in our analysis.

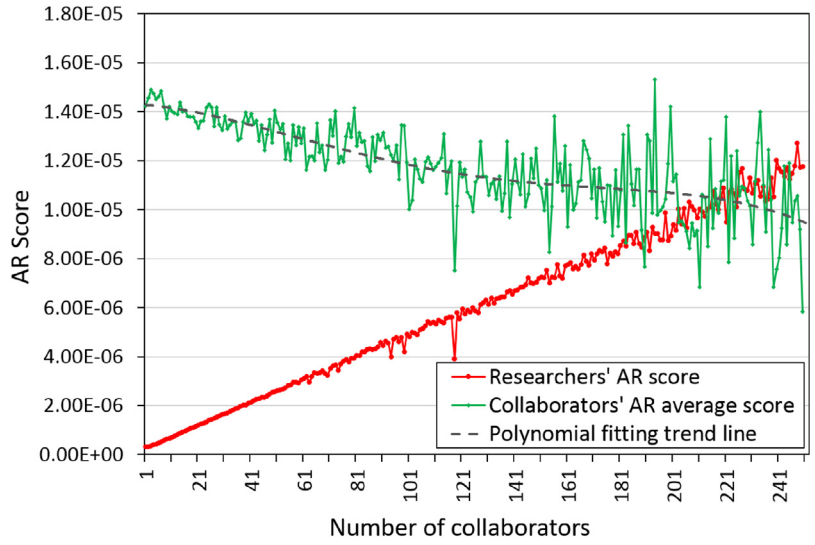

Fig. 3. Variation of *AR* score.

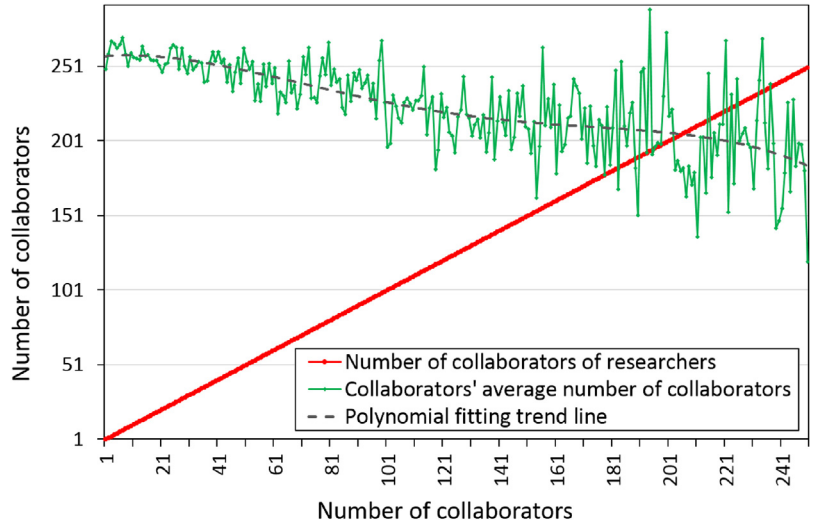

Fig. 4. Variation of number of collaborators.

4.2. Results

Fig. 2 shows the distribution of the number of researchers' collaborators. Most researchers have less than 30 collaborators, whereas the most common number of collaborators is three. However, some powerful researchers have more than 200 collaborators. The statistics show that the distribution of the number of researchers' collaborators does not fit a pure power-law distribution but is a long-tailed distribution. In the whole co-authorship network, the researchers who have little influence are still in the majority.

Corresponding to the three parts of Section 3, we will discuss the three aspects of our analysis results.

4.2.1. Friendship paradox validation

We measured the variation of researchers' *AR* score, number of collaborators, and number of publications, as well as these metrics for their collaborators, on average, to explore the friendship paradox. First, we built the collaboration network based on the DBLP dataset and calculated the *AR* score for each researcher according to Eq. (1). Similarly, the number of publications and collaborators was also counted in advance. Second, we averaged these three metrics for these researchers' collaborators. The results are shown in Figs. 3–5.

Fig. 3 shows that, as the number of collaborators increases, the *AR* score of researchers (red line) shows approximately linear growth. Researchers become more influential with increasing collaboration. We plotted a polynomial

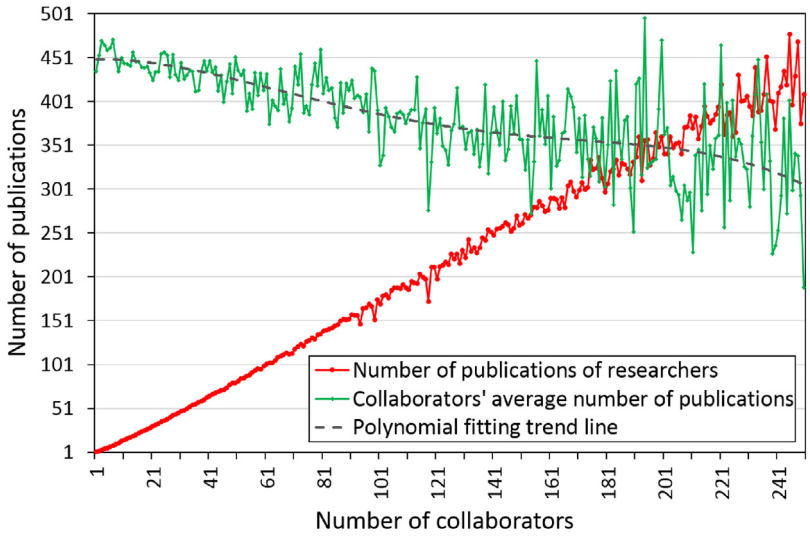

Fig. 5. Variation of number of publications.

fitting trend line (gray line) based on the green line, and the gray line gently declines. The fitting line follows the formula $y = -7 \times 10^{-15}x^4 + 3 \times 10^{-12}x^3 - 4 \times 10^{-10}x^2 - 9 \times 10^{-9}x + 10^{-5}$ with $R^2 = 0.5354$. Considering the average *AR* score of a researcher's collaborators, we can conclude that, although a researcher's rank is promoted by increased collaboration, the researcher's social circle changes, so that the average research rank of the collaborators gradually reduces.

The results clearly indicate that, when the scale of collaboration is not large, the average research rank of the collaborators is much higher than that of the researcher. This phenomenon verifies the generalized friendship paradox theory: "your friends are more prominent than you." The turning point of the generalized friendship paradox occurs when the number of the researcher's collaborators is large enough. When the researcher has more than 200 coauthor relationships with others, he/she will break the generalized friendship paradox and become more influential than the average for his/her social circle.

Figs. 4 and 5 show the variation in the number of collaborators and the number of publications, respectively, for researchers and their collaborators. As the number of collaborators increases, the number of publications (red line) of researchers shows approximately linear growth. With respect to the number of publications, researchers become more prolific with an increasing number of collaborators. We also plotted polynomial fitting trend lines (gray lines) for the collaborators' average number of collaborators and of publications (green lines). The fitting lines follow the formulas $y = -1 \times 10^{-7}x^4 + 7 \times 10^{-5}x^3 - 0.0101x^2 + 0.1694x + 257.6$ with $R^2 = 0.4839$ for number of collaborators, and $y = -2 \times 10^{-7}x^4 + 0.0001x^3 - 0.0159x^2 + 0.0828x + 449.13$ with $R^2 = 0.5066$ for number of publications. The two gray lines decline slowly, in a similar manner to the average *AR* scores of researchers' collaborators. This further demonstrates the conclusion: collaboration pushes researchers to break the generalized friendship paradox and makes researchers more prominent than others in their social circles.

4.2.2. Benefits of collaboration on a researcher's social circle

To measure the influence of collaboration on a researcher's social circle, we use the four metrics described in Section 3 to evaluate a researcher's social circle.

Before we find new collaborators of researchers in our dataset, we need to make sure that the influence is brought by new collaborators. Thus, we need firstly find common neighbors for each pair of co-authors before their first collaboration. For example, if *A* and *B* had common neighbors before their collaboration, *A* and *B* are new collaborators for each other. We need to find common neighbors of researchers *A* and *B* before their first collaboration.

In Fig. 6, the green line is the number of all collaborators and the red line represents the number of new collaborators. Fig. 6 shows that the number of new collaborators always increases monotonically. Moreover, the new collaborators introduced by original collaborators occupy a stable proportion of a researcher's social circle that cannot be ignored (almost 25%). Collaboration creates more opportunities for researchers to find new collaborators and allows researchers to have more collaborators. At the same time, collaboration broadens the research circles of scholars and expands the ego network of scholars. Fig. 7 shows that, as the number of collaborators increases, the proportion of new collaborators shows a rapid increase at first, followed by gentle growth. The value is close to 25% for researchers with more than 180 collaborators. We can conclude that a larger number of collaborators increase the probability of introducing new collaborators. The social circle of researchers will enter a positive cycle of growth with more collaborators. The analysis results of Fig. 7 verified the triadic closure theory in a certain respect. There is a high probability that co-authorship will create new collaborations for researchers. These new opportunities promote academic exchanges for scholars. Figs. 6 and 7 show that the number of new collaborators increases as the number of this researcher's collaborators increases. This social phenomenon explains the

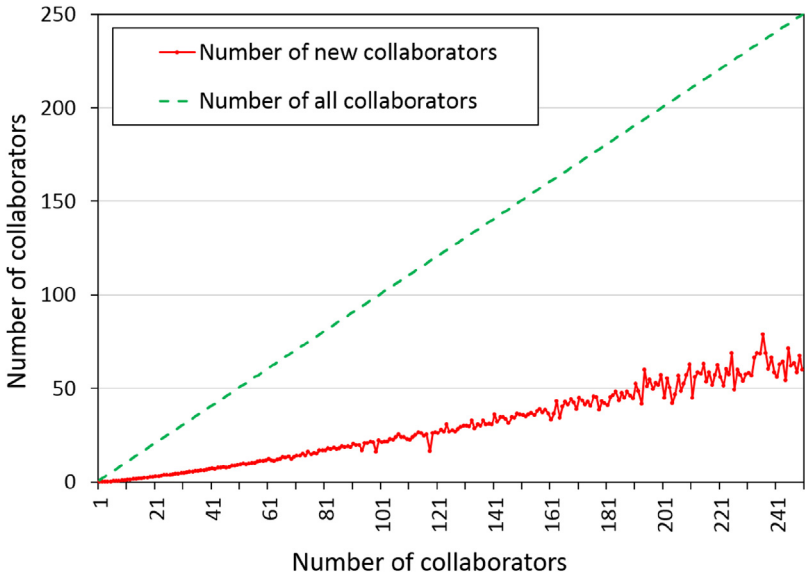

Fig. 6. Variation of number of new collaborators.

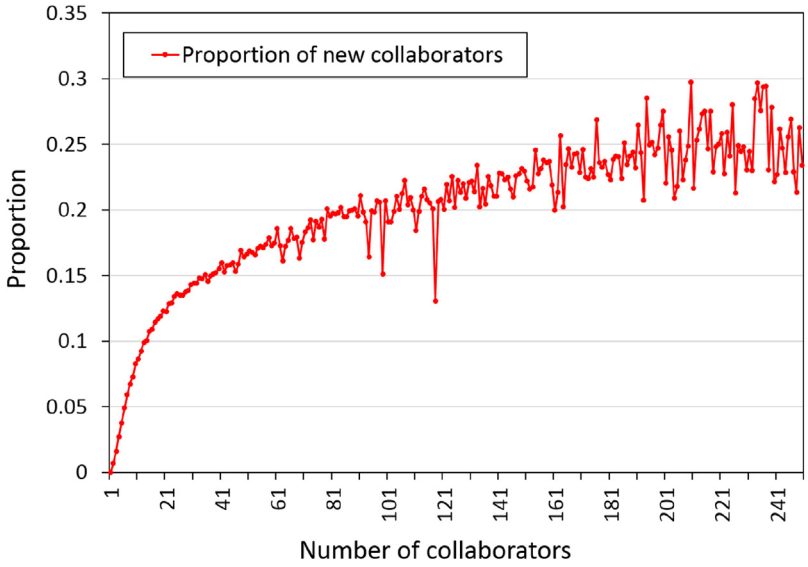

Fig. 7. Variation of proportion of new collaborators.

effect of preferential attachment in a certain respect: the probability that one researcher would like to collaborate with a researcher who has more collaborators is greater than with those with fewer collaborators. In Fig. 8, the number of potential collaborators also shows a remarkable growth rate as the number of collaborators increases. When a new researcher becomes a collaborator of a scholar, each collaborator of this researcher becomes a potential collaborator of the scholar. Therefore, collaboration indeed introduces many potential collaborators.

For the "Bacon" theory, Bacon is a central node in the network, and the Bacon Number is the social distance between other nodes and the Bacon node. Like "Bacon", there are stars who symbolize some domains in the co-authorship network. However, there are many research domains and central nodes. Moreover, to reduce the experimental error, we chose 100 top researchers as the "Bacon" nodes according to their *AR* scores. As shown in Fig. 9, the Bacon Number is the average social distance from researchers to these 100 "Bacons".

Fig. 9 shows the average, maximum, and minimum Bacon Number, respectively. We can see that, in general, the social distance in the co-authorship network is under 5.5, which accords with the "six degrees of separation" theory. As the number of collaborators increases, the average Bacon Number of researchers decreases from 5.36 to 3.5. This means that more collaboration can shorten the social distance between the researcher to the "Bacon" researchers (the representatively influ-

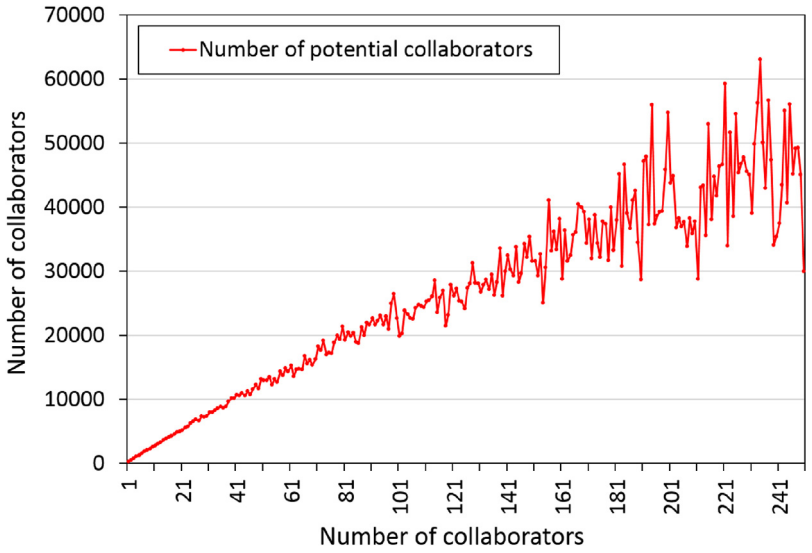

Fig. 8. Variation of number of potential collaborators.

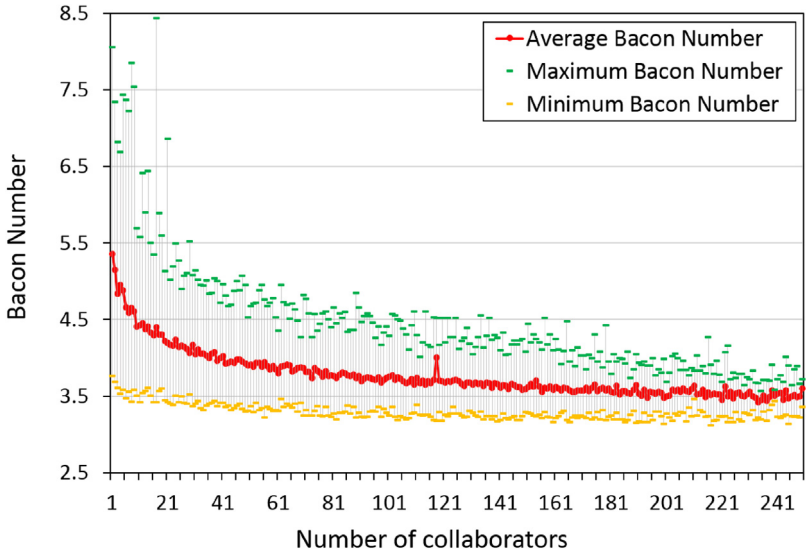

Fig. 9. Variation of Bacon Number.

ential researchers), and the relationship between scholars is closer. This in turn means that, as the number of collaborators increases, a researcher's social circle becomes more important and influential.

We can conclude that, based on triadic closure theory, collaboration will introduce new collaborators, broaden and enhance a researcher's social circle, and make it well-connected. As the social circle develops, the influence of researchers increases and researchers will become more popular and influential.

4.2.3. Variation of researcher's ego network

We demonstrate the influence of collaboration on a researcher's ego network, focusing on five aspects: eigenvector centrality, betweenness centrality, closeness centrality, clustering coefficient, and network density. Generally, with an increase in the number of collaborators, researchers become more influential and their positions become closer to the core position in the co-authorship network.

Fig. 10 shows the average, maximum, and minimum eigenvector centrality, respectively. The eigenvector centrality increases gradually with the increase in the number of collaborators. Moreover, when the number of collaborators is less than 180, the influence of researchers in the co-authorship network increases almost linearly with the number of collaborators. From Eq. (2), we know that eigenvector centrality is determined by the scores of neighbor nodes. This can explain why one researcher will achieve higher scores with more collaborators: his/her collaborators achieve high scores at the same time.

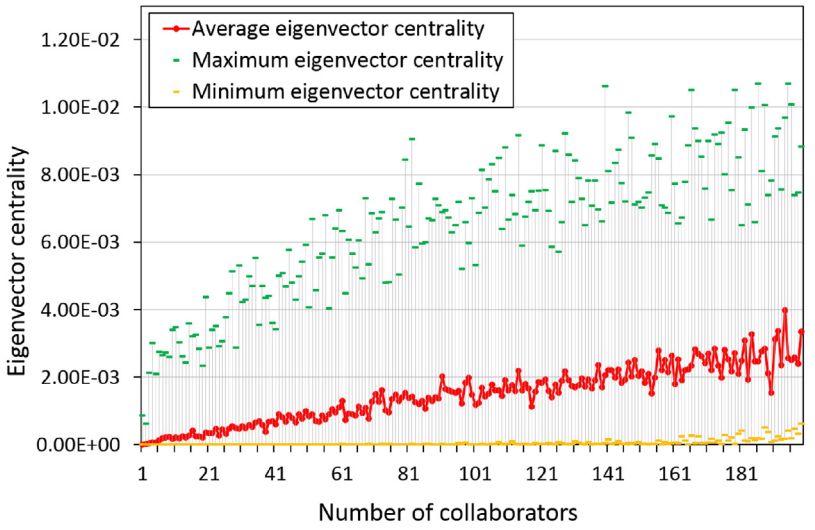

Fig. 10. Variation of eigenvector centrality.

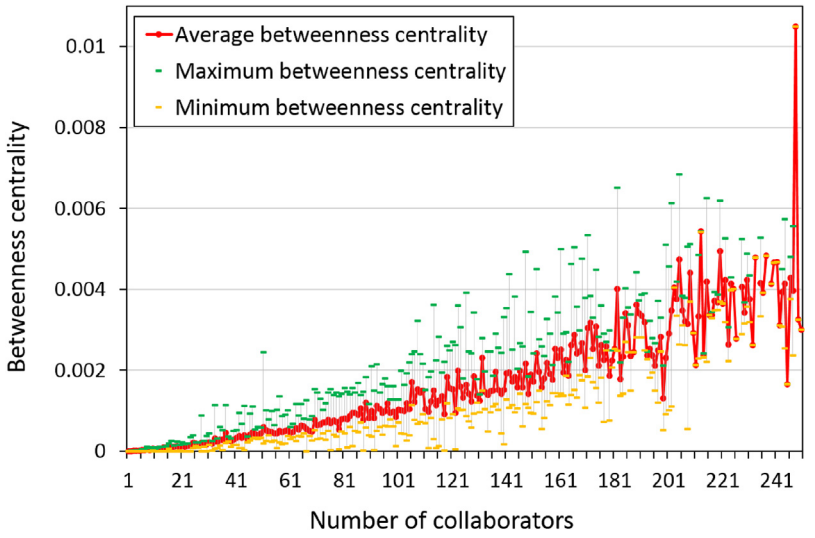

Fig. 11. Variation of betweenness centrality.

Fig. 11 shows the variation of betweenness centrality of researchers with an increasing number of collaborators. As the number of collaborators increases, the betweenness centrality shows a general increasing trend, but after the number of collaborators reaches 180, the value fluctuates a little. To some extent, we can conclude that the researchers play a more important role as intermediary with an increased number of collaborators.

Fig. 12 shows that the closeness centrality of researchers increases rapidly at first, then grows slowly, and shows a tendency toward stabilization at last, with an increasing number of collaborators. Moreover, the closeness centrality of researchers gradually stabilizes at around 0.24 when the number of collaborators exceeds 200. From the perspective of closeness centrality, researchers will become increasingly influential in terms of access to other researchers in the network, particularly when the number of collaborators is less than 50, but their influence will be stable later. According to Eq. (4), when a researcher has more collaborators, the distance from all researchers in the co-authorship network to this researcher will be shorter, so this researcher will tend to be at the core of the network.

Fig. 13 shows that the clustering coefficient generally decreases as the number of collaborators increases. We can find that the clustering coefficient of researchers in the co-authorship network ranges from 0 to 1. It is evident that the clustering coefficient is greater when there are fewer nodes, which demonstrates that the clustering coefficient is relatively high in the preliminary ego network. However, with an increasing number of collaborators, the clustering coefficient becomes lower, because the aggregation of the researchers will disperse and the coverage of researchers will be larger when they have more collaborators, which translates to a larger denominator in Eq. (5).

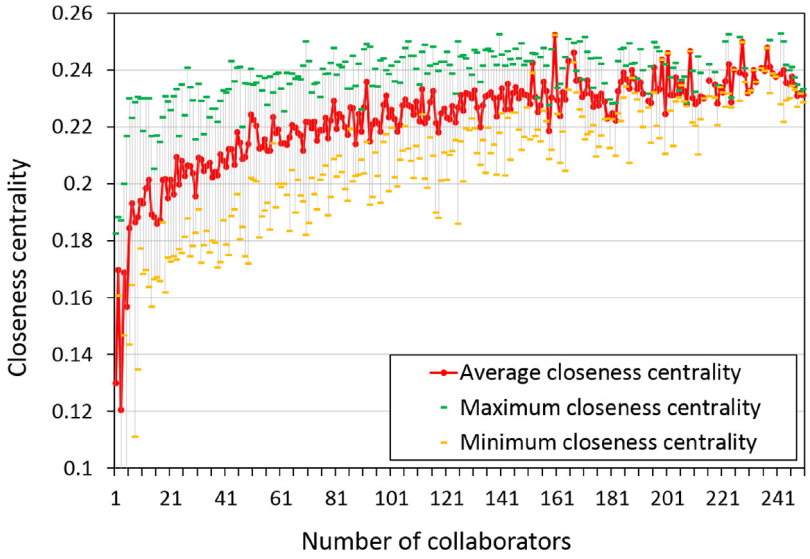

Fig. 12. Variation of closeness centrality.

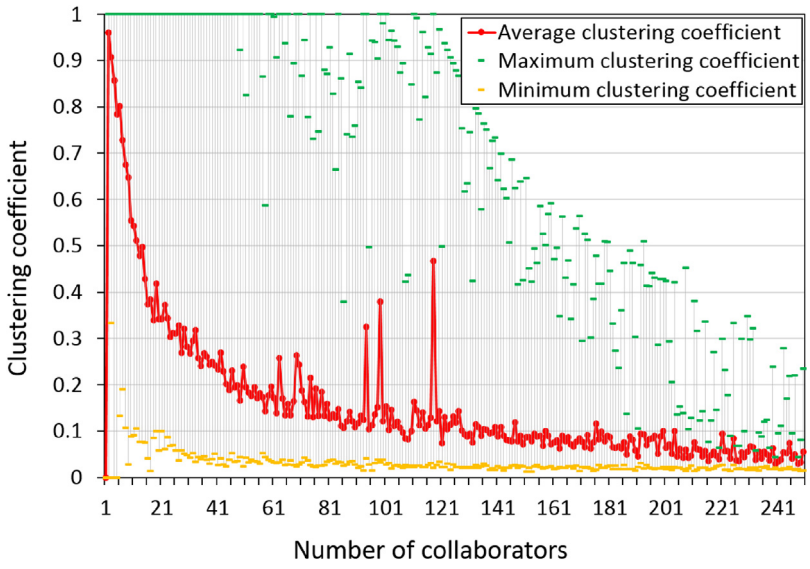

Fig. 13. Variation of clustering coefficient.

Fig. 14 shows that the trend of network density is similar to that of the clustering coefficient. With an increase in the number of collaborators, communication between researchers will be more frequent.

We observe that scholars become closer to the core in the co-authorship network and more influential with an increase in the number of collaborators. Therefore, we can conclude that collaboration makes researchers more likely to be closer to these excellent researchers, from the perspective of network structure. Collaboration does therefore promote the position of researchers in their co-authorship network.

5. Conclusion

Studying the relationships within scientific collaborations with respect to the impact of researchers is an evergreen topic. Many studies have focused on recommendation techniques to strengthen research collaboration or measure the rank and impact of researchers; however, we wanted to explore the mechanism behind these two phenomena. In this paper, we mainly explored how collaboration brings researchers more impact, through data analytics, focusing on three main aspects: impact of the researchers, benefits gained by researchers, and changes to the structure of the whole co-authorship network. We used social analytical methods to quantify the collaboration and the impact of researchers, with respect to network structure

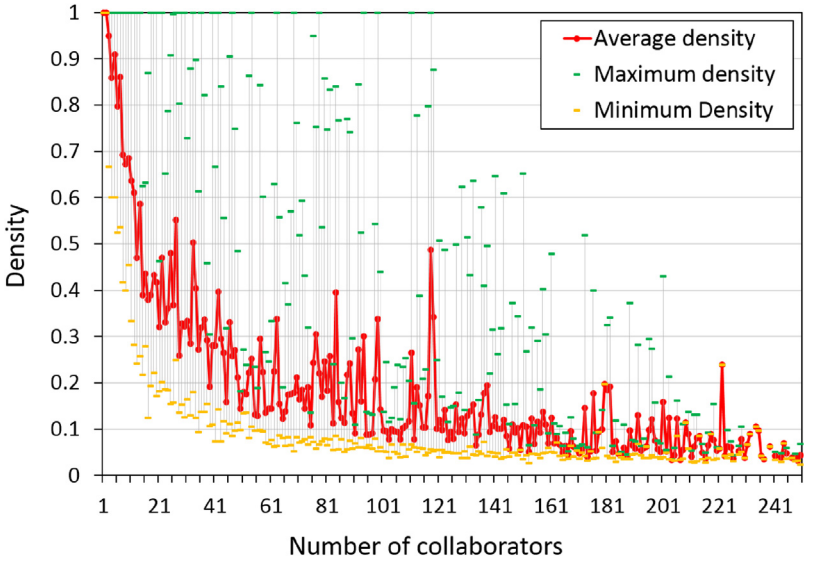

Fig. 14. Variation of density.

and individual characteristics. By introducing the "six degrees of Bacon Number" theory, generalized friendship paradox, and triadic closure theory to interpret our analysis, we observed the benefits that collaboration brings to researchers. The results show that collaboration improves research. Increasing the number of collaborations results in more papers and more collaborators and increases the influence of researchers. Moreover, collaborations broaden and enhance a researcher's social circle and make it well-connected. The whole co-authorship network will become increasingly complex and the relationship between researchers will be more stable. Our findings reveal the hidden patterns behind collaboration. The insights obtained may shed light on studying scientific collaboration from a new angle.

Authors contribution

Xiangjie Kong: Conceived and designed the analysis, performed the analysis, wrote the paper.
Mengyi Mao: Contributed data or analysis tools, performed the analysis, wrote the paper.
Huizhen Jiang: Collected the data, performed the analysis.
Shuo Yu: Collected the data, performed the analysis.
Liangtian Wan: Performed the analysis.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China under Grants (61801076, 61872054), and was supported by the Fundamental Research Funds for the Central Universities under Grants (DUT18JC09, DUT19LAB23).

References

Amjad, T., Ding, Y., Xu, J., Zhang, C., Daud, A., Tang, J., & Song, M. (2017). Standing on the shoulders of giants. *Journal of Informetrics*, *11*, 307–323.
Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2012). Four degrees of separation. *in: Proceedings of the 4th Annual ACM Web Science Conference, ACM.*, 33–42.
Barabási, A., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*, 590–614.
Barr, J., & Gunawardena, A. (2012). Classroom salon: a tool for social collaboration. *in: Proceedings of the 43rd ACM technical symposium on Computer Science Education, ACM.*, 197–202.
Benevenuto, F., Laender, A. H. F., & Alves, B. L. (2016). The h-index paradox: your coauthors have a higher h-index than you do. *Scientometrics*, *106*, 469–474.
Castro, R. D., & Grossman, J. W. (1999). Famous trails to paul erdos. *Mathematical Intelligencer*, *21*, 51–53.
Coccia, M., & Wang, L. (2016). Evolution and convergence of the patterns of international scientific collaboration. *Proceedings of the National Academy of Sciences*, *113*, 2057–2061.
Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J. P., Sanchez, Á., et al. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, *214*, 325–346.
van Dijk, D., Manor, O., & Carey, L. B. (2014). Publication metrics and success on the academic job market. *Current Biology*, *24*, R516–R517.
Dunaiski, M., Geldenhuys, J., & Visser, W. (2018). Author ranking evaluation at scale. *Journal of Informetrics*, *12*, 679–702.
Ebadi, A., & Schiffauerova, A. (2015). How to become an important player in scientific collaboration networks? *Journal of Informetrics*, *9*, 809–825.
Eom, Y. H., & Jo, H. H. (2014). Generalized friendship paradox in complex networks: The case of scientific collaboration. *Scientific Reports*, 4.

Feld, S. L. (1991). Why your friends have more friends than you do. *American Journal of Sociology*, *96*, 1464–1477.
Fotouhi, B., Momeni, N., & Rabbat, M. G. (2015). Generalized friendship paradox: An analytical approach. *Lecture Notes in Computer Science*, *8852*, 339.
Ghobadi, S. (2015). What drives knowledge sharing in software development teams: A literature review and classification framework. *Information & Management*, *52*, 82–97.
Granovetter, M. S. (1977). The strength of weak ties. *Social Networks*, *78*, 347–367.
Grund, T. (2014). Why your friends are more important and special than you think. *Sociological Science*, *1*, 128–140.
Hodas, N.O., Kooti, F., Lerman, K., 2013. Friendship paradox redux: Your friends are more interesting than you. arXiv preprint arXiv:1304.3480.
Hsiao, K. J., Kulesza, A., & Hero, A. (2014). *Social collaborative retrieval*. pp. 293–302.
Kim, J., & Diesner, J. (2017). Over-time measurement of triadic closure in coauthorship networks. *Social Network Analysis & Mining*, *7*, 9.
Kong, X., Jiang, H., Wang, W., Bekele, T. M., Xu, Z., & Wang, M. (2017). Exploring dynamic research interest and academic influence for scientific collaborator recommendation. *Scientometrics*, *113*, 369–385.
Kong, X., Jiang, H., Yang, Z., Xu, Z., Xia, F., & Tolba, A. (2016). Exploiting publication contents and collaboration networks for collaborator recommendation. *PloS ONE*, *11*, e0148492.
Kong, X., Mao, M., Wang, W., Liu, J., & Xu, B. (2018). VOPRec: Vector representation learning of papers with text information and structural identity for recommendation. *IEEE Transactions on Emerging Topics in Computing*, http://dx.doi.org/10.1109/TETC.2018.2830698
Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social studies of science*, *35*, 673–702.
Lerman, K., Yan, X., & Wu, X. Z. (2016). The "majority illusion" in social networks. *PLOS ONE*, *11*, 1–13.
Leskovec, J., & Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. *in: Proceedings of the 17th international conference on World Wide Web, ACM.*, 915–924.
Liu, J., Kong, X., Xia, F., Bai, X., Wang, L., Qing, Q., & Lee, I. (2018). Artificial intelligence in the 21st century. *IEEE Access*, http://dx.doi.org/10.1109/ACCESS.2018.2819688
Liu, Y., Venkatanathan, J., Goncalves, J., Karapanos, E., & Kostakos, V. (2014). Modeling what friendship patterns on facebook reveal about personality and social capital. *ACM Transactions on Computer-Human Interaction (TOCHI) 21*, *17*(1-17), 20.
Mavin, A. (2015). I'll tell you what i want, what i really, really want: An industry perspective on the effective application of research in projects. *in: Conducting Empirical Studies in Industry (CESI), 2015 IEEE/ACM 3rd International Workshop on, IEEE.*, 34.
Momeni, N., & Rabbat, M. (2016). Qualities and inequalities in online social networks through the lens of the generalized friendship paradox. *PLOS ONE*, *11*, 1–17.
Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, *64*, 025102.
Nykl, M., Campr, M., & Jezek, K. (2015). Author ranking based on personalized pagerank. *Journal of Informetrics*, *9*, 777–799.
Opsahl, T. (2013). Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, *35*, 159–167.
Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: bringing order to the web*.
Rafailidis, D., & Crestani, F. (2016). Collaborative ranking with social relationships for top-n recommendations. *in: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM.*, 785–788.
Sarigol, E., Pfitzner, R., Scholtes, I., Garas, A., & Schweitzer, F. (2014). Predicting scientific success based on coauthorship networks. *EPJ Data Science*, *3*, 9.
Schoenebeck, G. (2013). Potential networks, contagious communities, and understanding social network structure. *in: Proceedings of the 22nd international conference on World Wide Web, ACM.*, 1123–1132.
Staiano, J., Lepri, B., Aharony, N., Pianesi, F., Sebe, N., & Pentland, A. (2012). Friends don't lie: inferring personality traits from social network structure. *in: Proceedings of the 2012 ACM conference on ubiquitous computing, ACM.*, 321–330.
Thelwall, M. (2018). Do females create higher impact research? scopus citations and mendeley readers for articles from five countries. *Journal of Informetrics*, *12*, 1031–1041.
Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the facebook social graph. *Computer Science*.
Wang, W., Yu, S., Bekele, T. M., Kong, X., & Xia, F. (2017). Scientific collaboration patterns vary with scholars' academic ages. *Scientometrics*, *112*, 1–15.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440–442.
Weitzel, L., Quaresma, P., & de Oliveira, J. P. M. (2012). Measuring node importance on twitter microblogging, in: Proceedings of the 2nd International Conference on Web Intelligence. *Mining and Semantics, ACM.*, *11*(1-11), 7.
Xia, F., Chen, Z., Wang, W., Li, J., & Yang, L. T. (2014). MVCWalker: Random walk-based most valuable collaborators recommendation exploiting academic factors. *IEEE Transactions on Emerging Topics in Computing*, *2*, 364–375.
Xia, F., Wang, W., Bekele, T. M., & Liu, H. (2017). Big scholarly data: A survey. *IEEE Transactions on Big Data*, *3*, 18–35.
Zhang, C., Bu, Y., & Ding, Y. (2016). Understanding scientific collaboration from the perspective of collaborators and their network structures. *IConference 2016 Proceedings*.
Zhang, C., Bu, Y., Ding, Y., & Xu, J. (2017). Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *Journal of the Association for Information Science & Technology*, *69*, 72–86.
Zhong, E., Fan, W., Zhu, Y., & Yang, Q. (2013). Modeling the dynamics of composite social networks. *in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.*, 937–945.
Zuckerman, E. W., & Jost, J. T. (2001). What makes you think you're so popular? self-evaluation maintenance and the subjective side of the "friendship paradox". *Social Psychology Quarterly*, *64*, 207–223.